

Faculty of Engineering and Information Technology
University of Technology Sydney

**Online Media Information
Enhancement by Fusing Multimodal
Media Data**

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Lu Zhang

March 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Lu Zhang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date:

09/03/2021

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my principal supervisor, Associate Prof. Jian Zhang, for his professional guidance and help. He never gave up on me even when I stuck in my progress. I deeply appreciate my co-supervisor Jingsong Xu, for his help and guidance in both research and living in Sydney. He used to spend a lot of time to help me revise my paper and discuss my research problem. Also, I would like to thank Reader Jialie Shen for his guidance and advice in our collaborative research. Without his help, the research will not be completed successfully.

Besides, I want to thank my friend Yongshun Gong. Thanks for his patience and kindness. He always gave me generous help. I want to thank Zhibin Li. Thank him for his company and encouragement. Sometimes he was very strict with me and I believe he must have his reasons. I might not be able to finish my study without them. The time spent with them is my precious memory. I also want to thank myself. Thank you for being in a foreign country and supporting yourself with braveness and strength on countless nights. May the mountain remain the mountain. May the sea remain the sea. May we still be us.

Finally and most essentially, I would like to thank my parents. Without them, nothing would have any value.

Lu Zhang

March 2021 @ UTS

Contents

Certificate of Original Authorship	i
Acknowledgment	ii
List of Figures	vii
List of Tables	xi
List of Publications	xii
Abstract	xiii
 Chapter 1 Introduction	 1
1.1 Background	1
1.2 Applications	3
1.3 Research Question	5
1.3.1 Multimodal marketing intention understanding	5
1.3.2 Multimodal travel information matching	7
1.4 Research Challenges	9
1.4.1 Heterogeneity between multiple modalities	10
1.4.2 Different contributions from multiple modalities to specific task	10
1.4.3 Correlation and independence between multiple modalities	12
1.4.4 Challenges based on specific application scenarios	13
1.4.5 Summary	14
1.5 Research Contributions	14
1.6 Thesis Structure	16

Chapter 2 Literature Review	18
2.1 Summary	18
2.2 Methods based on Levels of Fusion	18
2.2.1 Early fusion	19
2.2.2 Late fusion	20
2.2.3 Hybrid fusion	21
2.3 Methods based on Specific Models	22
2.3.1 Kernel-based methods	22
2.3.2 Correlation-based methods	23
2.3.3 Bayesian inference methods	24
2.3.4 Multimodal Deep Learning	24
2.4 Related works of multimodal marketing intention understanding	25
2.4.1 Human intention analysis	25
2.4.2 Text-based marketing content detection	26
2.4.3 Graph Neural Networks	27
2.4.4 Topic Model	27
2.5 Related works of travel information understanding	29
2.5.1 Travel information analysis	29
2.5.2 Multiple Kernel K-means	30
 Chapter 3 Two-branch Cross-graph Fusion Strategy for Mul-	
timodal Marketing Intention Detection	31
3.1 Introduction	31
3.2 Main Contribution	35
3.3 Two-branch Cross-graph Accumulative Fusion Method for Mul-	
timodal Marketing Intention Detection	37
3.3.1 Notations	37
3.3.2 Problem Formulation	37
3.3.3 Methodology	38
3.3.4 Experiment	40
3.4 Two-branch Cross-graph Assignment Fusion Method for Mul-	
timodal Marketing Intention Detection	44

3.4.1	Methodology	44
3.4.2	Experiment	46
3.5	Conclusion	47

Chapter 4 Supervised Multimodal Document Informed Neural Autoregressive Distribution Estimator for Multimodal Marketing Intention Analysis 49

4.1	Introduction	49
4.2	Main Contribution	52
4.3	Supervised Multimodal Document Informed Neural Autoregressive Distribution Estimator	52
4.3.1	Definitions and Problem Formulation	54
4.3.2	Methodology	58
4.4	Applications and Datasets	62
4.4.1	Applications	62
4.4.2	Dataset Construction	63
4.5	Experiment	66
4.5.1	Experimental Setup	67
4.5.2	Evaluation Metrics and Baselines	68
4.5.3	Results on DS1, DS2 and DS3	68
4.5.4	Case Studies	73
4.6	Conclusions	74

Chapter 5 Hybrid Multiple Kernel K-means for Multimodal-based Travel Information Enhancement 76

5.1	Introduction	76
5.2	Main Contribution	80
5.3	Hybrid Multiple Kernel K-means Model	82
5.3.1	Problem definition	83
5.3.2	Text embedding	83
5.3.3	Text-image joint embedding	85
5.3.4	Multiple Kernel Clustering	87

5.4	Experiments	91
5.4.1	Dataset construction	91
5.4.2	Experimental Setup	93
5.4.3	Ablation study	95
5.4.4	Comparison with baselines	97
5.4.5	Image/text-to-text kernel evaluation	99
5.4.6	Visualisation	100
5.5	Conclusion	101
Chapter 6	Conclusions and Future Work	102
6.1	Conclusion	102
6.2	Future Work	104
	Bibliography	105

List of Figures

1.1	An example of the multimodal data on media platform. . . .	2
1.2	An example of a marketing advertorial.	5
1.3	(a) A travelogue example from web forum TripAdvisor. (b) A landscape image example from online platform Flickr.	8
1.4	Examples advertorials: (a) An advertorial example in which commercial intention is expressed only by the text. (b) An advertorial example in which commercial intention is conveyed only by the image.	11
1.5	An example in which the commercial intention is embedded in both images and text.	12
1.6	The illustration of the thesis structure.	17
2.1	The illustration of the early fusion strategy.	19
2.2	The illustration of the late fusion strategy.	20
2.3	The illustration of the hybrid fusion strategy.	21
3.1	The structure of the proposed two-branch cross-graph accumulative fusion method. Images and texts are represented as graph structures and then are fed into two GCNs. Then the graph center is calculated and crosswise accumulated. Graph pooling layer is followed after fusion.	35
3.2	The structure of the proposed two-branch cross-graph assignment fusion method.	36
3.3	Network architecture.	39

3.4	An example in the dataset. From the texts, it is easy to identify the marketing intention. However from only images, the topic is blurry.	43
4.1	Advertorial examples from media platforms. The above one is a piece of normal social news that shares knowledge about running. The middle one shows a biscuit advertorial with marketing intention embedded mainly in images. The below one shows a clothing advertorial with marketing intention embedded mainly in text.	50
4.2	Framework of the proposed multimodal based marketing intention analysis system.	53
4.3	Logic sequence of three proposed research questions.	63
4.4	This Figure shows the statistics of DS1 regarding the number of images per piece of social news. The horizontal axis represents the number of images per piece of social news. The vertical axis represents the number of pieces of social news.	64
4.5	This Figure shows the statistics of DS1 regarding the number of sentences per piece of social news. The horizontal axis represents the number of sentences per piece of social news. The vertical axis represents the number of pieces of social news.	64
4.6	This Figure shows the statistics of DS2/3 regarding the number of images per piece of social news. The horizontal axis represents the number of images per piece of social news. The vertical axis represents the number of pieces of social news.	66
4.7	This Figure shows the statistics of DS2/3 regarding the number of sentences per piece of social news. The horizontal axis represents the number of sentences per piece of social news. The vertical axis represents the number of pieces of social news.	66
4.8	Detailed performance comparison of accuracy over 5 topics on DS2.	70

4.9	Detailed performance comparison of accuracy over 3 levels on DS3.	71
4.10	Examples for marketing intention extent analysis. (a) ‘weak’ extent; (b) ‘medium’ extent; (c) ‘strong’ extent.	72
4.11	Examples of advertorials that were unsuccessfully predicted. .	74
5.1	(a) Travelogue data examples. (b) Image data example. Each image has three textual components as descriptive information: ‘title’, ‘tag’ and ‘description’.	77
5.2	Flowchart of the proposed method. Three kinds of data (travelogues from TripAdvisor, images and the descriptive texts from Flickr) are used as inputs. Two different methods are adopted to embed texts and images. A multiple kernel clustering method is proposed to fuse the embedding features for multi-source travelogue-image matching.	81
5.3	Text-image joint embedding. Both images and texts are used in this process. First, images and texts are encoded by different methods into separate subspaces. Then the encoded vectors are fed into a two-branch network to project image and text features into a common latent subspace. Based on the correlations in the common subspace, the kernel matrices can be built.	85
5.4	Unsupervised clustering for image and travelogue matching. The blue circles represent images. The orange squares represent travelogues. Matching images will be recommended to each travelogue: 1) if the number of images in the same class with the travelogue is enough, these images will be recommended based on their similarity scores (See the green arrowed lines in the graph). 2) If there is no more image in the same class to be recommended, images in the nearest class will be considered (See the red arrowed lines in the graph).	91

5.5	One example of travelogue enhancement. On the top is the travelogue information. Under the travelogue, there are the matching results from TFIDF, Word2Vec, Doc2Vec, and the proposed method. The underlines show the words that have semantically similar words between the travelogue and the descriptive texts. The hexagons show the words that appear both in the travelogue and the annotations.	98
-----	--	----

List of Tables

3.1	ImageGraph and TextGraph are constructed and random split into training, validation and testing sets.	40
3.2	Comparison with five baselines.	42
3.3	Advertorial detection accuracy (%) and F1-score (%) comparison. The highest accuracy and F1-score are highlighted. . . .	46
3.4	Results of ablation study. The highest accuracy and F1-score are highlighted.	47
4.1	Notations and their definitions.	57
4.2	Statistics of datasets DS2 and DS3.	67
4.3	Accuracy and F1-score Comparison.	69
4.4	Results of Ablation Study.	71
5.1	MAP Comparison on Different Kernel Combination by Gaussian Kernel.	92
5.2	MAP Comparison on Different Kernel Combination by Linear Kernel.	92
5.3	Comparison Under Different Regularisation Terms.	93
5.4	Performance Comparison with Baselines.	94
5.5	Comparison of Using Different Kernel Groups in Terms of MAP.	94

List of Publications

Papers Accepted and Published

- **Lu Zhang**, Jingsong Xu, Yongshun Gong, Litao Yu, Jian Zhang, and Jialie Shen, *Unsupervised Image and Text Fusion for Travel Information Enhancement*, IEEE Transactions on Multimedia, 2021.
- **Lu Zhang**, Jialie Shen, Jian Zhang, Jingsong Xu, Zhibin Li, Yazhou Yao, and Litao Yu, *Multimodal Marketing Intent Analysis for Effective Targeted Advertising*, IEEE Transactions on Multimedia, 2021.
- **Lu Zhang**, Jian Zhang, Jialie Shen, Jingsong Xu, Zhibin Li, and Litao Yu, *Incorporating Multimodal Cues for Advertorial Discovery*, International IEEE Conference on Multimedia and Expo, 2021.
- **Lu Zhang**, Jian Zhang, Zhibin Li, and Jingsong Xu, *Towards Better Graph Representation: Two-Branch Collaborative Graph Neural Networks for Multimodal Marketing Intention Detection*, International IEEE Conference on Multimedia and Expo, 2020.
- **Lu Zhang**, Jingsong Xu, Jian Zhang, and Yongshun Gong, *Information Enhancement for Travelogues via a Hybrid Clustering Model*, Digital Image Computing: Techniques and Applications, 2018.

Abstract

With the rapid improvement of human’s ability to store and compute big data, more and more people are used to sharing and acquiring multimodal information on media platforms. The explosion of multimodal information is paving the way for many innovative applications based on the mining and analysis of big multimodal data. This thesis focuses on two applications: how to explore multimodal cues to recognise and analyse marketing intentions embedded in normal social news and how to mine correlative multimodal information for travel information understanding and recommendation.

Multimodal fusion is to join the associated features or integrate the intermediate decisions from two or more modalities in order to perform a prediction or an analysis for specific tasks. The key challenge behind this topic is the inherent heterogeneity of multimodal data leading to semantic gaps between them. Given a large amount of multimodal data, the main factor in establishing an effective model for specific tasks is the deep mining of inter-correlations between multiple modalities. Moreover, the intra-correlations within the same modality are also important for semantic inferring. This thesis develops several supervised and unsupervised methods to overcome these challenges.

To infer the commercial intentions embedded in multimodal content of social news, this thesis proposes a supervised Neural Autoregressive Topic Model which utilizes both textual words, visual words, and label information to learn discriminative representations for specific tasks. Moreover, this thesis introduces a supervised Two-branch Collaborative Graph Neural Net-

work, which uses a pre-defined fusion strategy to integrate news text and news images. A novel Multimodal Advertorial Discovery Model is proposed by adopting a Cross-graph Fusion strategy to achieve a soft assignment to incorporate images and text and generate comprehensive multimodal representations. For travelogues enhancement, this thesis introduces an unsupervised Hybrid Multiple Kernel K-means method to embed multimodal for matching travelogues and images. To verify the effectiveness of these proposed methods, experiments are conducted on real-life datasets from online media platforms.